REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-00-

00 36

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1 Jun 99	FINAL - 1 MAR 98 - 28 FEB 99

4. TITLE AND SUBTITLE
Personal Computer Cluster for Theoretical Studies of Gas Phase Elementary Reactions

5. FUNDING NUMBERS
F49620-98-1-0345

6. AUTHOR(S)
Professor Keiji Morokuma

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Emory University
Department of Chemistry
1515 Pierce Drive
Atlanta GA 30322

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NL
801 N Randolph St., Rm 732
Arlington VA 22203-1977

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Three graduate students, a postdoctoral fellow, and Emerson Center staff and workstation expert and the PI of this project formed a team and work together very hard. The team did extensive literature studies (mainly visiting many web sites), detailed discussions on technical aspects had jpricing studies(via internet and fax) before ordering any hardware and software. In particular, Dr. Szilagyi served as the team captain and put often overheated discussions into reality. Mr. Khoroshun has been the system manager of our PCcluster, and spend hundreds of hours in front of the Master Console in istalled, tested, and integrated new hardware and software into the system. It is quite obious that a few/several devoted graduate student and/or postdoctoral fellows are essential for any success of this kind of PC project.

14. SUBJECT TERMS
PC Cluster, hardware, software

15. NUMBER OF PAGES
11

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UU	UU	UU	

Final Technical Report

to the Air Force Office of Scientific Research
Bolling Air Force Base, 110 Duncan Ave.,
Washington, D. C. 20322-8080
Program Director: Dr. Mike Berman, Theoretical Chemistry

Submitted: June 1, 1999

The Principal Investigator:
Keiji Morokuma
Department of Chemistry, Emory University,
1515 Pierce Dr., Atlanta, GA 30322
Phone: (404)-727-2180
E-mail: morokuma@emory.edu

AFOSR Grant Number: F49620-98-1-0345 (DURIP)
Project Period: Mar. 1, 1998 - Feb. 28, 1999

Title: Personal Computer Cluster for Theoretical Studies of Gas Phase Elementary Reactions

1. Introduction

The goal of the present project is to construct a cluster of personal computers (PCs) to provide relatively inexpensive but high-performing computational capability that is needed for execution of our AFOSR-supported research (F49620-98-1-0063). There are still relatively a small number of PC-based clusters with more than 20 machines constructed and used in individual research groups in university environment. While the modest present project is under way, other larger scale efforts were also making progress elsewhere. For instance, at the end of 1998, a 5000-CPU Intel-based supercomputer at Sandia National Laboratory was among the top ten fastest supercomputers in the world.

Computational resources at the Morokuma Group before the present installation were six IBM RISC 6000 cluster running AIX operating system and two-year old sixteen dual PentiumPro PCs running Linux operating system (we call Generation 1, **G1** PCs). Using our experience with this **G1**, the present project is to develop a PC cluster specifically suited for purposes of computational quantum chemistry.

At the completion of the present one-year Instrumental grant, we have constructed successfully a 49 unit, 98 processor PC cluster, consisting of 33 newly purchased dual processor PCs, with additional (existing) 16 dual processor PCs integrated. The system is operating very stably with nearly 100% up-time and expected speed and is used very heavily for research supported by the AFOSR grant (F49620-98-1-0063).

This final technical report describes the details of the current implementation of the PC cluster. Three major issues: a) hardware selection, b) software adaptation, and c) integration of the cluster were to be addressed separately.

2. The current implementation

All the major aspects of the PC-cluster purchased and constructed with the present grant are presented on Scheme 1. The photos of some of PCs (G3, G2 and G1, as discussed below) are in Figures 1-3. Details of the present project will be discussed in the following subsection.

At the beginning, I would like to emphasize that three graduate students (Mr. Dmitri V. Khoroshun, Ms. Zhiwei Liu and Mr. Alexey L. Kaledin), a postdoctoral fellow (Dr. Robert Szilagyi), an Emerson Center staff and workstation expert (Dr. Stephan Irle) and the PI of this project (Keiji Morokuma) formed a team and worked together very hard. The team did extensive literature studies (mainly visiting many web sites), detailed discussions on technical aspects and pricing studies (via internet and fax) before ordering any hardware and software. In particular, Dr. Szilagyi served as the team captain and put often overheated discussions into reality. Mr. Khoroshun has been the system manager of our PC cluster, and spent hundreds of hours in front of the Master Console in installed, tested, and integrated new hardware and software into the system. It is quite obvious that a few/several devoted graduate student and/or postdoctoral fellows are essential for any success of this kind of PC project.

A. The PC hardware

With the present grant, we purchased 12 Generation-2 (**G2**) *production* PCs in August 1998 and 18 Generation-3 (**G3**) *production* PCs in February 1999 along with 3 *pilot* PCs (**T1**, **T2** and **T3**) as the major component of the present PC cluster based on Pentium-dual processor machines. We give the details of the configuration of individual machines in Appendix 1. They are all in full operation at the end of the grant period and are used very heavily (nearly 100% utility) for our AFOSR-supported research. With the purchase of a total of 33 PCs with 66 CPUs

from this grant, integrated with our existing cluster of 16 Generation-1 (**G1**) PCs with 32 CPUs, our entire cluster now consists of 49 PCs with 98 CPUs.

We have spent a substantial time in carefully designing the PC cluster system with the best cost performance for the computational needs of our group. Pentium CPU was selected over the DEC alpha CPU, because the latter had no stable compiler that runs under Linux system; otherwise, one had to purchase an expensive alternative operating system. No monitor will be purchased for any PCs, since these were to be connected with network and used without direct user access. The operation of all the individual machines is to be monitored with a Master Console with a monitor we already owned.

For the purchase of the PC hardware, the galloping development of faster components proved to necessitate very careful consideration and testing. In one case, directly after the two test PCs (**T1** and **T2**), based on the Intel LX440 chipset (66MHz frequency bus), were successfully tested, a newer BX440 chipset (100MHz bus) emerged as an alternative with a better price/performance ratio. Consequently, 12 **G2** PCs were purchased without rigorous test of a prototype, which leaded to a hardware instability problem due to the incompatibility with the motherboard (SuperMicro P6DBE). The only solution found so far was to switch **G2** PCs into 66MHz main bus frequency mode, while increasing the multiplication factor for the internal CPU bus from 4 to 5. Effectively, 400(internal bus)/100(main bus) MHz computers currently work in 333/66MHz mode, losing about 20% computational power but gaining the expected complete stability of the system. The subsequent purchase of 18 **G3** PCs (utilizing a better motherboard) was based on careful tests of exactly the same prototype (test PC **T3**), and no stability problem was experienced for already 3 months of uninterrupted full load production work.

Selection of the hard drives was an important part of the design. While cheaper 7,200 rpm Ultra-ATA IDE hard drives were used for **G2** production PCs, newer **G3** machines, equipped with 10,000 rpm UltraWide SCSI-2 hard drives, were shown to perform considerably superior for disk-extensive benchmark jobs. We expect that the overall I/O speed is at least 2.5 times larger for the SCSI-2 **G3** PCs. For the data server (see Scheme 1), combination of two 9GB 10,000 rpm UW SCSI-2 drives (software and immediate user data) together with a 23 GB 5,400 rpm SCSI-2 drive (for storage of secondary user data) satisfied the needs of the users so far.

We purchased a backup tape drive each for data and user servers (see Scheme 1). Additionally, we purchased a CD-RW device, which is interfaced to the user server and has been used for transferring user data to the CD ROM media.

Accelerated video cards and 21" monitors were purchased for both data and user servers, and are used with the Linux X-Window system, for which many useful visualization and support programs are available. The purchase of the Master Console devices (currently with 28 keyboard/mouse/monitor inputs) was very useful, by permitting the system manager to monitor virtually all PCs without having to switch the cables (see Scheme 1).

Finally, a vital hardware-related issue of our PC cluster worth mentioning is the Fast Ethernet (100Mb) network setup, as shown in Appendix 2. One 8-port Fast Switch acts as a local 100Mb backbone for our cluster, while being connected to the Emory University 10Mb Ethernet, allowing therefore for communication with the outside world. The quality of the current network setup, best judged by performance of distributed parallel applications, still remains to be tested, as our parallel application at present is restricted in running two CPUs in parallel in one machine and does not involve communication between two machines connected by the network. Many

models of 100Mb switches, while being an obvious improvement of 10Mb Ethernet, are known to be of a quality insufficient for network-based parallel applications. Complicating the issue even further, the Linux network subsystem setup plays an important role for such applications. Additional tests would definitely be required in order to judge the quality of our PC cluster's network.

B. The software

The fastest growing Linux (a POSIX-compliant UNIX) operating system (RedHat distribution, version 5.2) is currently used for all computers in the cluster. The well-known stability, performance, excellent software and troubleshooting support and documentation of this easy-to-use free operation system made it a natural choice fir our PC cluster. The available SMP (Symmetric Multi Processor) feature of Linux allows for fully independent usage of both CPUs in a dual box without any additional effort. However, our experience with **G2** PCs confirmed the fact that stability of the SMP implementation is particularly sensitive to selection of the hardware. Further, a standard SVR4 UNIX IPC (InterProcess Communication) feature of Linux was found to be in a development stage; the latest 2.2 family of kernels still does not allow for allocation of more than 128MB of shared memory. Nevertheless, the current implementation of IPC found its use in small-memory **G1** PCs of our cluster, where a single job utilizes both CPUs of each box running in parallel/shared memory mode, with the speedup factor of 1.95-2.00. The compiler of we purchased, Portland Group F77, has proven to be superior (by about 30%) to the freely available, generic (but poorly optimized) analogs, g77 and f2c/gcc.

Timely notifications and immediate patches of the security problems, appreciated by all Linux users, were extremely helpful. Nevertheless, prior to activation of *TCP wrappers* (a very "soft" and easy-to-use analog of a firewall), our system has been compromised twice. Introduction of TCP wrappers, in combination with other security tools such as tripwire, has successfully protected our cluster from observable general or directed attacks for already about half a year.

An extremely important part of the software is the queuing system. We used a freely available, simple Distributed Queuing System, DQS, version 3.2. We developed efficient submitting scripts as well as a particular queuing system ideological setup that would balance different needs of different users.

C. The integration of the system

The our PC cluster may be described as a *weakly bound cluster*, meaning essentially that the production nodes work mostly in sequential, as contrasted to parallel, mode. The resulting setup is very similar to that of very popular clusters of IBM RS6000 workstations. The large (about 20) number of users, as well as relatively modest computational demands of each particular job, allowed for such a setup, in which the main feature is a large number of small jobs submitted by many users. Consequently, it was very important for us to integrate the system properly so that it complies with the demands of a multi-user, remote access environment.

Some of the hardware features were designed specifically for the integration purposes. The tape and CD-RW backup devices described above reduce the amount of the necessary disk space. Separation of the data and user servers allows to dedicate the data server for performing a quite extensive task of being an NFS server for almost 50 clients. The NFS performance is quite acceptable due to the Fast 100Mb Ethernet. Finally, 4 additional remote terminals (4 iMacs purchased from the present grant) were necessary to accommodate the growing number of users.

In some cases, the complexity of chemical problems requires color presentations or multimedia documentation. Along this line, we purchased a cost-effective Ink Jet color printer capably to handle super-A3 size poster sheets. A Mac computer was arranged to be dedicated for multimedia applications and live presentations with 19" screen monitor.

We finally decided not to purchase an SGI graphic workstation, proposed in the original grant proposal. The main reason for this is that many graphic packages are supported by Mac OS and Linux environment, and introducing a different platform was judged to be counterproductive. Instead, we decided to use the available funds for purchase of more **G3** PCs and user-support peripherals, such as back-up tape drive, CD-RW and larger hard disk storage, discussed above.

The software and system setup also follows the necessity of supporting a large number of users with divergent requirements. Important details are the user disk quotas on the data partitions and a fairly set up queuing system sensitive to varying from user to user computational demands. Separation of the data and user servers, with the crucial data server and all the production machines being *completely closed* to the outside world access for security reasons, became possible due to a properly designed system integration setup. Finally, many quantum chemical packages (such as Gaussian94/98/99, MOLPRO96/98, GAMESS96, Hondo95, Turbomol) and many corresponding integration scripts, as well as general-purpose packages (such as Molden3 and LaTeX) are made available to support diverse needs of the users.

3. Future improvement and expansion plans

Among our future improvement and expansion plans are:

A. Implementation of different distributed parallelization techniques for various quantum chemistry packages (Gaussian + Linda, GAMESS + Global Arrays, Turbomole + PVM), involving porting the software, tuning of the network subsystem setup, and possibly modernization of the network hardware.

B. Improvement of the user diversity possibilities, involving updates of the queuing system setup, submission scripts, compilation of various software as well as improving the porting of the currently present software.

C. Improving the security of the system, vital in the currently aggressively hostile Internet environment. The ultimate goal is implementation of the so-called IP-masquerading technique, which would make all the working nodes, as well as the data server, completely invisible and inaccessible to the outer network behind a Linux PC-based router.

D. Installation of up-to-date improved system elements, the most crucial of which is the system kernel. Implementations of both SMP and IPC are currently under way in the Linux community.

4. Appendix

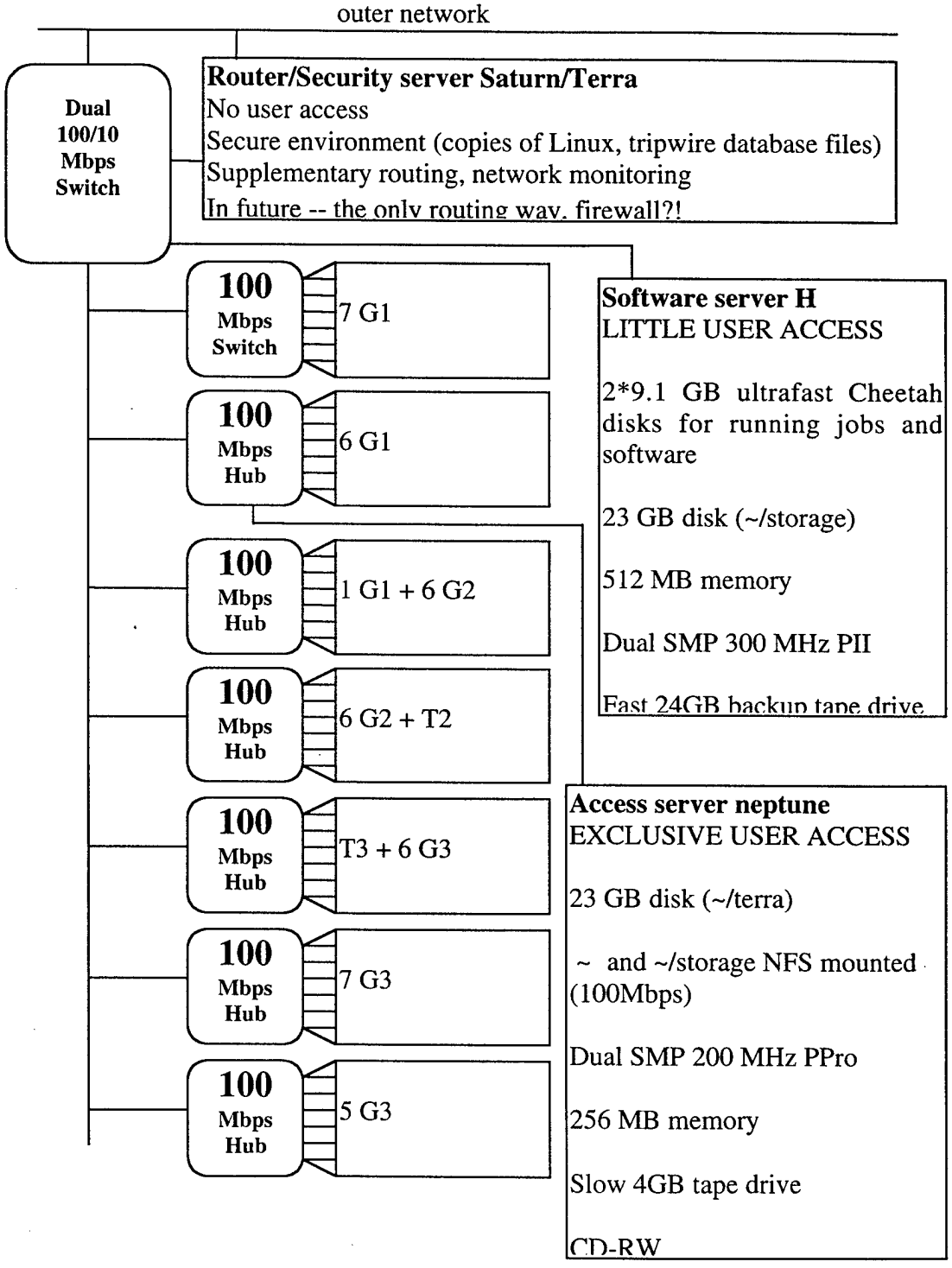
Appendix 1. Details of the configurations of the purchased *production* PCs (**G2** and **G3**) as well as *pilot* PCs (**T1**, **T2** and **T3**).

	T1	**T2**	12x **G2**	**T3**	18x **G3**
Motherbord	P6DLS (SuperMicro)	P6DLE (SuperMicro)	P6DBE (SuperMicro)	Asus P2B-DS	Asus P2B-DS
CPU	2x Intel 300 MHz PII	2x Intel 266 MHz PII	2x Intel 400 MHz PII	2x Intel 450 MHz PII	2x Intel 450 MHz PII
Memory	4x 10ns 128MB	4x	4x 8ns 128MB	4x 8ns 128MB (Siemens)	4x 8ns 128/256MB (Siemens)[a)]
Hard drive	2x 9.1 GB 10,000 rpm UW SCSI-2 (Seagate), coolers. 23.2 GB 5,400 rpm SCSI-2, (Seagate) cooler	9.1 GB 7,200 rpm Ultra IDE (Seagate)	9.1 GB 7,200 rpm Ultra IDE (Seagate)	9.1 GB 10,000 rpm UW SCSI-2 (IBM), cooler	9.1 GB 10,000 rpm UW SCSI-2 (IBM), cooler
Network card	KNE 100TX 100Mb (Kingston)	KNE 100TX 100Mb (Kingston)	KNE 100TX 100Mb (Kingston)	KNE 100TX 100Mb (Kingston)	KNE 100TX 100Mb (Kingston)
Case, video card, floppy drive	Full tower case, 2MB Trident video card, 1.44MB Floppy drive	Mid tower case, 2MB Trident video card, 1.44MB Floppy drive	Mid tower case, 2MB Trident video card, 1.44MB Floppy drive	Full tower case, 2MB Trident video card, 1.44MB Floppy drive	Full tower case, 2MB Trident video card, 1.44MB Floppy drive

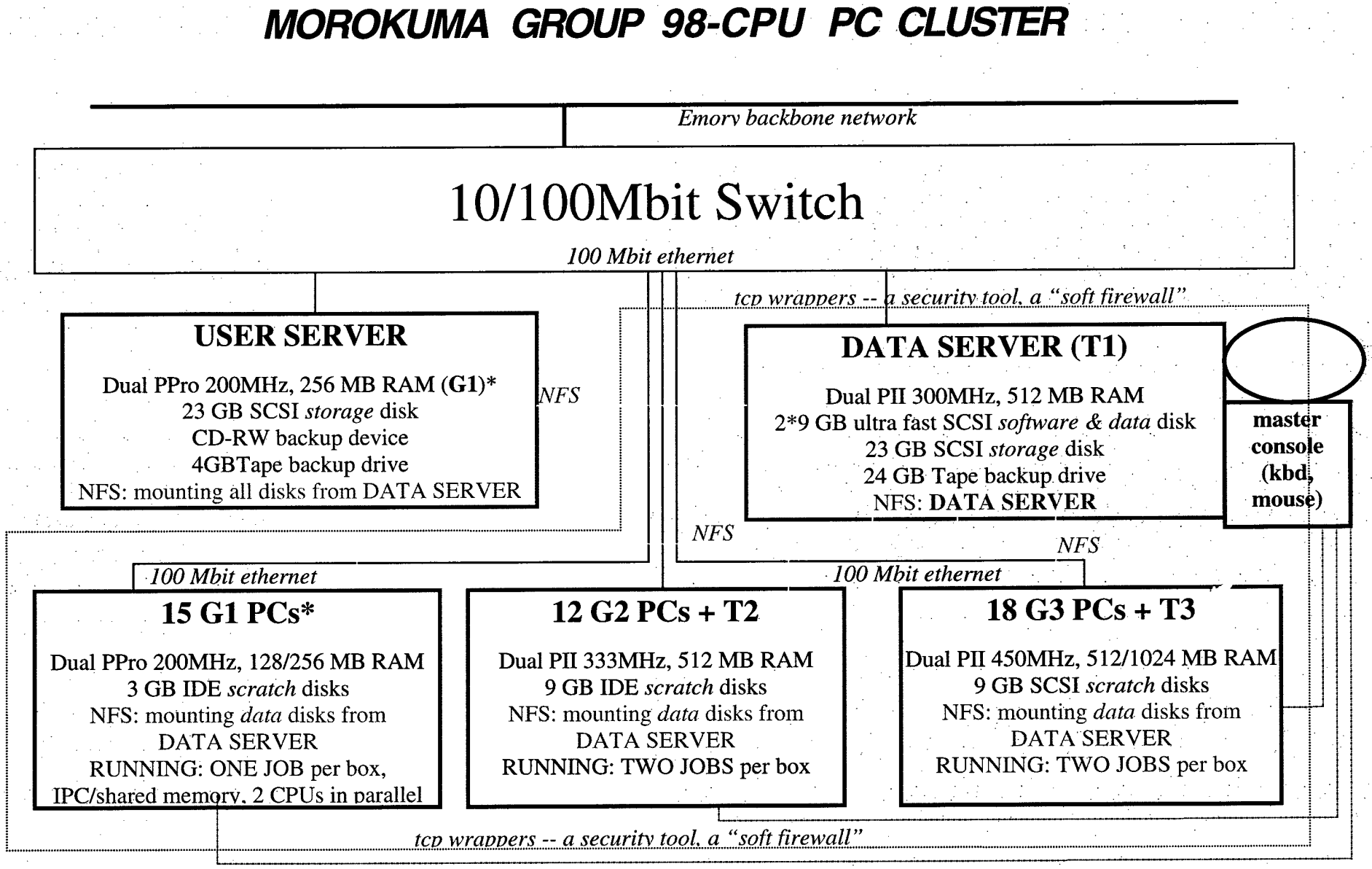
a) 10 of **G3** PCs have 4*128MB RAM setup, 4 have 2*256MB and the other 4 have 4*256MB.

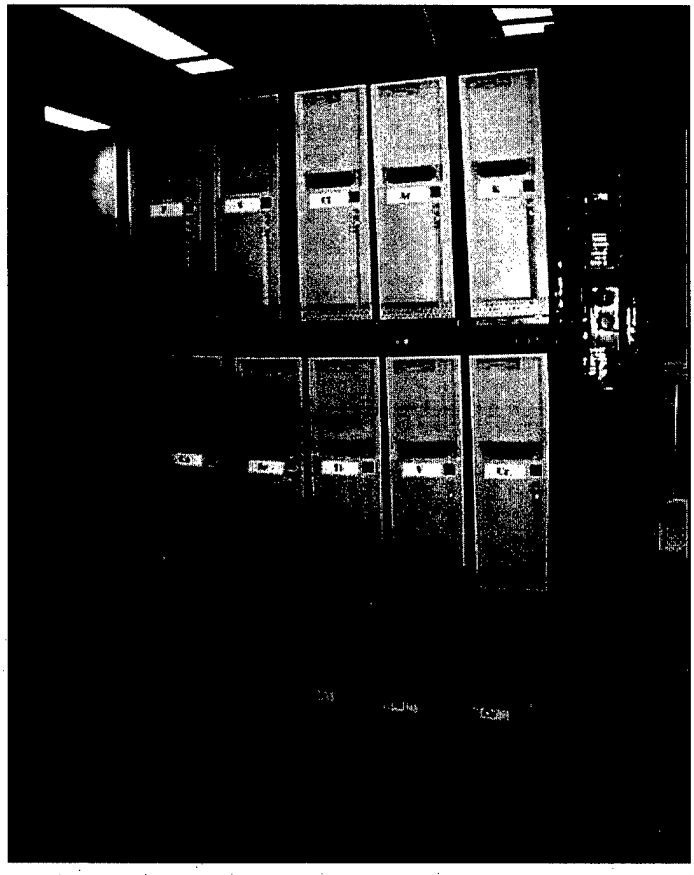
Appendix 2. Network components and configuration.

One 8-port 10/100 Mb NetGear FS4008E Switch: a backbone of the present PC cluster system and an uplink to University network/Internet.
One FastEthernet 100Mb 8-port Netgear FS108 Switch:an uplink of the production nodes and for distributed parallel computations.
Six FastEthernet 100Mb 8-port NetGear FE108 Hubs: Uplink of individual PCs

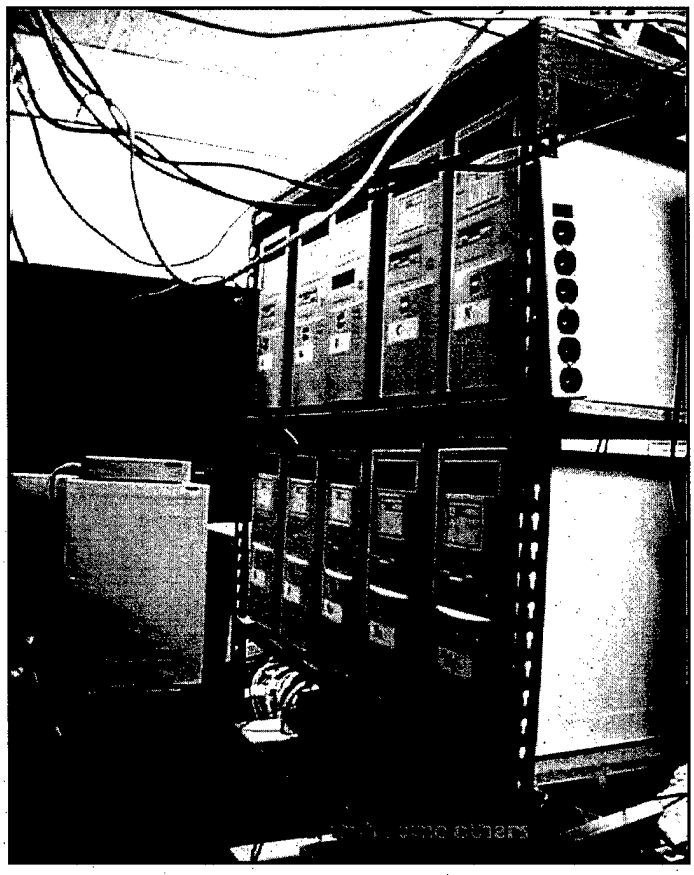


Scheme 1. A general scheme of the Morokuma Group PC cluster setup. (Hardware existing prior to this grant is marked with *)





PC4.jpg



PC8.jpg